**Fraud Detection**

Problem Statement: To detect fraudulent insurance claims made by patients and hospitals.

Datasets:

1. Benificiary dataset – Dataset contains patient demographics, disease diagnosed and the claim details of all the patients (inpatient and outpatient).
2. Provider dataset – Claim details like start and end date, patient details like amount paid, discharge date and hospital details like attended physician and other physician.
3. Train Dataset – Contains two variables, provider\_id and potential fraud(True/False)
4. Test Dataset – Contains provider\_id and the values should be predicted as true or false. True – Indicates the claims are fraudulent, False – Indicates the claims are not fraudulent.

Data Prep:

Checking for missing values – There were missing values in the age column and it is imputed by taking the max date of death and calculating the age.

Creating new variables from existing variables

1. Patient deceased or not
2. Admission date
3. Discharge date
4. Number of days admitted

Merging Beneficiary, Provider and train dataset.

Creating dummy variables for categorical variables.

Exploratory Data Analysis:

Potential non fraud claims are more in number than the potential fraud claims.

PRV51459 – Huge number of fraudulent claims.

Feature Engineering:

Grouping different features and checking if the grouped patterns lead to fraudulent behaviours.

Model:

Standardize the data and perform train test split.

Logistic regression model is built to find if the claims are fraudulent or not.

Dependent Variable – Provider ID

If the predicted probability of fraud is < 0.7, the claims are not Fraudulent else the claims are fraudulent.